Обучение машинного перевода без параллельных текстов*

Скиднов Е. А. Бахтеев*[1] *О. Ю. Стрижов*[2] *В. В.

[1]Московский физико-технический институт

[2]Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН

Данная задача посвящена исследованию алгоритма обучения машинного перевода без параллельных предложений. Использование параллельных текстов для задачи машинного перевода требует слишком большой базы предложений всех переводимых языков, что является ресурсоемкой задачей для некоторых пар непохожих языков. Особенностью исследуемого алгоритма является то, что для перевода используется кодировние и декодирование текста во внутреннем представлении. Данный алгоритм использует единую модель нейронной сети Seq2Seq для перевода с одного языка на другой и обратно. Цель данного исследования заключается в том, чтобы сделать вектора скрытых пространств этих двух моделей как можно более похожими. Для демонстрации работоспособности метода будет использован вычислительный эксперимент машинного перевода между двумя похожими языками: русским и украинским.

Ключевые слова: *машинный перевод, нейросеть, Seq2Seq.*

1 Введение

Решается задача оптимизации системы машинного перевода без использования параллельных предложений. Так как для некоторых пар языков получение таких пар предложений, а также и само обучение является ресурсоемкой задачей [Koehn .Koehn .2007, KoehnKoehn2009].

Существует ряд подходов к построению систем машинного перевода. Предлагается использовать рекуррентные нейронные сети с короткой и долгой памятью и нейронные сети, в которых реализовано механизм внимания (attention). В данном методе используются нейронные сети, которые осуществляют перевод в два этапа. Такой метода называется Seq2Seq [Weiss, Chorowski, Jaitly, Wu ChenWeiss .2017].

Данная работа посвящена последнему методу последовательного перевода. Предлагается с помощью первой рекуррентной нейронной сети, основанной на долгой памяти перевести входящую последовательность в вектор, а с помощью второй перевести этот вектор в выходную последовательность на нужном нам языке [Cho, Van Merriënboer, Bahdanau BengioCho .2014]. Наша модель оптимизируется таким образом, чтобы скрытые пространства для векторов предложений двух языков совпадали. Данный метод позволяет гораздо быстрее обучить нейронную сеть переводу с одного языка на другой, в связи с использованием ей предыдущего опыта и наличию у нее памяти и внимания.

Эксперименты и анализ качества предложенного метода проводится на паре языков "русский-украинский"с помощью алгоритма BLEU(Bilingual evaluation understudy) для проверки качества текста [Papineni, Roukos, Ward ZhuPapineni .2002], переведенного с одного языка на другой.

*Работа выполнена при финансовой поддержке РФФИ, проект №00-00-00000. Научный руководитель: Стрижов В. В. Авторы: А.В. Грабовой, О.Ю. Бахтеев, В.В. Стрижов, Eric Gaussier, координатор Малиновский Г.С. Консультант: Бахтеев О. Ю.

Литература

[Cho, Van Merriënboer, Bahdanau BengioCho .2014] cho2014propertiesCho, K., Van Merriënboer, B., Bahdanau, D. Bengio, Y. 2014. On the properties of neural machine translation: Encoder-decoder approaches On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259.

[KoehnKoehn2009] koehn2009statisticalKoehn, P. 2009. Statistical machine translation Statistical machine translation. Cambridge University Press.

[Koehn .Koehn .2007] koehn2007mosesKoehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N.others 2007. Moses: Open source toolkit for statistical machine translation Moses: Open source toolkit for statistical machine translation. Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions Proceedings of the 45th annual meeting of the acl on interactive poster and demonstration sessions (177–180).

[Papineni, Roukos, Ward ZhuPapineni .2002] papineni2002bleuPapineni, K., Roukos, S., Ward, T. Zhu, WJ. 2002. BLEU: a method for automatic evaluation of machine translation Bleu: a method for automatic evaluation of machine translation. Proceedings of the 40th annual meeting on association for computational linguistics Proceedings of the 40th annual meeting on association for computational linguistics (311–318).

[Weiss, Chorowski, Jaitly, Wu ChenWeiss .2017] weiss2017sequenceWeiss, RJ., Chorowski, J., Jaitly, N., Wu, Y. Chen, Z. 2017. Sequence-to-sequence models can directly translate foreign speech Sequence-to-sequence models can directly translate foreign speech. arXiv preprint arXiv:1703.08581.

Поступила в редакцию